# Learning to Buy Time: A Data-Driven Model For Avoiding Silence While Task-Related Information Cannot Yet Be Presented

**Soledad López Gambino**
Bielefeld University
Universitätsstraße 25
33615 Bielefeld
Germany

**Sina Zarrieß**
Bielefeld University
Universitätsstraße 25
33615 Bielefeld
Germany

**Casey Kennington**
Boise State University
1910 University Dr.
Boise, Idaho
USA

**David Schlangen**
Bielefeld University
Universitätsstraße 25
33615 Bielefeld
Germany

m.lopez_gambino,sina.zarriess,david.schlangen@uni-bielefeld.de
casey.kennington@cs.boisestate.edu

## Abstract

Current dialogue systems typically do not explicitly manage time. Where attempts are made at rectifying this, this is often done with a focus on turn-taking, aiming at making the system take the turn more quickly and naturally. Here we look at another, related phenomenon: what to do when the turn has been taken, but the expected task-related content cannot yet be produced. We implemented a system that can produce what we call "time buying" acts whenever it needs to bridge time until it can present a task-level reply (in our case, flight information). The range of acts and, crucially, the sequencing of these acts (including their temporal placement) are learned from an existing corpus in which such situations were created on purpose. We evaluate this system by letting participants interact with it as well as with two baseline systems: one that only produces one type of act (namely explicitly asking the user to wait) at regular intervals, and another one that produces the full range of acts, but sequenced randomly. We find that participants rate the full system as more human-like than the other systems and that they also report enjoying interacting with it more. We conclude that "buying time" in a natural fashion is possible and beneficial for interaction quality, but only if sequencing constraints found in natural data are reproduced.

## 1 Introduction

Consider the following interaction between a caller who wants to book an airline ticket and a travel agent:

| | | | |
|---|---|---|---|
| (1) | Caller: | I'd like to book a flight from Aachen to Zurich for next Monday. | [1] |
| | Agent: | Of course. Aachen to Zurich, | [2] |
| | | *[agent inputs information to find possible flights]* | [3] |
| | | uhm... | [4] |
| | | I'm starting to get some results here | [5] |
| | | just one more moment | [6] |
| | | Ok, so there is a flight on Monday at ... | [7] |

Here, the agent produces utterances that are not strictly task-relevant, but still seem to fulfill an interaction management function.[1] In general, dialogue participants seem to try to avoid longer pauses in dialogue (Lundholm Fors, 2015; Jefferson, 1989), since delays are often interpreted as signs of a problem (Levinson, 1983; Kohtz and Niebuhr, 2017). This is especially true if the speakers are not co-located (as in the example above) and lack information from other modalities such as gaze and facial expression.

Speakers use a variety of resources to bridge time in dialogue, including fillers (e.g. line 4 in (1)) and explicit requests for waiting (e.g., line 6), but also other kinds of utterances such as echoing the interlocutor's words (e.g., line 2), committing themselves to the task (line 2) or conveying the state of the information (e.g. line 5) (Clark and Tree, 2002; López Gambino et al., 2017).

In this paper, we explore modeling this kind of time-buying behavior in a spoken dialogue system, and we evaluate perceived naturalness and enjoyment when human participants interact with it. The language

[1] This particular interaction is constructed for clarity, but similar ones are attested for example by López Gambino et al. (2017); see below.

| | | | |
|---|---|---|---|
| **SYSTEM**: | *Reiseinformationssystem DSG-Bielefeld. Danke, dass Sie uns nochmals anrufen. Was kann ich für Sie tun?* | **SYSTEM**: | 'Travel Information System DSG-Bielefeld. Thank you for calling us again. How may I help you?' |
| **CALLER**: | *Hallo. Gibt es einen Flug mit dem Startflughafen Frankfurt und dem Zielflughafen Sydney am 3. August vormittags?* | **CALLER**: | 'Hello. Is there a flight with departure airport Frankfurt and destination airport Sydney on August 3, in the morning?' |
| **SYSTEM**: | *Mm-hm, gut. Die Flüge werden noch gesucht. Nach Sydney (...) einen kleinen Moment, bitte (...)*<br>*Ich warte noch auf die Liste, die Flüge kommen langsam rein (...)*<br>*Ich habe einen passenden Flug gefunden. Ich sende Ihnen die Daten per Email. Vielen Dank.* | **SYSTEM**: | 'Mm-hm, okay. The search for flights is still in progress. To Sydney (...) one moment, please (...)<br>I'm waiting for the list, the flights are appearing slowly (...)<br>I've found a matching flight. I'll send you the information by email. Thank you very much.' |
| **CALLER**: | *Danke.* | **CALLER**: | Thank you. |
| **SYSTEM:** | *Auf Wiederhören.* | **SYSTEM**: | Goodbye. |

Table 1: Example interaction between system and participant, from the data collected in the experiment. Original in German on the left, English translation on the right.

of the system is German. The utterances used to bridge time are inspired by those found in a corpus of human-human dialogues ((López Gambino et al., 2017), and Section 3.1.3 below), and the system also considers information on how these utterances are sequenced in human-human data. (See Table 1 for an example of an interaction with the system.)

To evaluate the system, we had participants interact with it and with two baseline systems. The first one bridges the gap between the user's request and presentation of a result by explicitly asking the user to wait. The second system uses the same utterances as the one based on human behavior but selects them randomly, without considering any sequencing information. After each dialogue, participants were asked to rate the system with which they had just interacted. Our system was rated as more human-like and more enjoyable to interact with than the other systems. Additionally, it was perceived as capable of finding a result in a more appropriate amount of time than the system which used explicit requests to wait, although the actual time elapsed before announcing a result was the same for all three systems. Below we describe the system and the evaluation, after looking at related work.

## 2 Related work

Previous studies in the field of automatic systems as well as customer satisfaction have focused on comparing strategies which can be applied during long waiting periods. Some of the strategies tested are playing music, or providing information about waiting time, place in the queue, or choice of listening alternatives (Tom et al., 1997; Antonides et al., 2002; Munichor and Rafaeli, 2007). One reported finding is that telephone systems which fill long gaps are perceived more positively by humans than those which remain silent until information can be presented (López Gambino et al., 2018; Tom et al., 1997), which is not surprising given humans' dislike of long pauses in dialogue (see Section 1). On the other hand, the conclusion that subjective perception of elapsed time depends (at least partly) on what the subjects hear in the meantime (derived from the results presented in 4.1) has been somewhat more contested, since previous literature presents evidence in its favor (Hirsch et al., 1950; Antonides et al., 2002) as well as against it (Tom et al., 1997; Munichor and Rafaeli, 2007).

More generally, the work presented here can be seen as part of current efforts on *incremental dialogue processing* (Skantze and Hjalmarsson, 2010; Schlangen and Skantze, 2011; Buß and Schlangen, 2010). This paradigm enables the development of systems which can manage time by strategically planning (and re-planning) the production of utterances and their timing. Such strategic decisions can depend on the internal state of the system (e.g. a system which is still in the process of generating an information utterance and produces a filler to cover the pause) or on external considerations (such as a system which reformulates an already planned utterance due to a recent change in the environment). One such system is described in (Skantze and Hjalmarsson, 2010). The system bridges the gap before information presentation either through fillers (e.g. *eh*) or by playing canned beginnings of utterances such as *It costs...* or *Here is a...*, which the system then completes with content synthesized online. Similarly, Baumann and Schlangen (2013) tested an incremental system which also uses open-ended utterances that are extended

| Action | Category | Example |
|---|---|---|
| PRODUCE GROUNDING UTTERANCE | acknowledgment | C: I want to fly to Bristol.<br>A: *Okay.* |
| | echoing | C: I'm looking for a flight to Izmir at the beginning of August.<br>A: *A flight to Izmir, beginning of August,* let me see... |
| | commitment | *Let's have a look.* |
| PRODUCE INFORMATION STATE UTTERANCE | agent/system state | *The search for flights is still in progress.* |
| | temporary non-availability | *Until now I haven't found any morning flights.*<br>*So far I only see evening flights.* |
| | wait request | *Please hold on.* |
| | availability | *We have a few choices to offer you.* |

Table 2: Actions and utterance categories

as new information comes in. In addition, this system introduces hesitations to compensate for long pauses resulting from overcommitment. Hesitations have also been employed by Betz et al. (2017) as a means for recovering the user's attention when it deviates away from the system. Another incremental system which reacts to events in its surroundings was presented by Buschmeier et al. (2012). The system reacted to noise interruptions in the environment by pausing its speech and re-generating the interrupted chunk once the noise had stopped. Stent (1999) presented a system which bridges gaps by inserting fillers or utterances such as *wait a minute*. Finally, Tsai et al. (2018) developed a movie recommender dialogue system which fills the silent time before information presentation by uttering a general statement, such as *I think this movie fits your tastes*. To the best our knowledge, the issue of modeling "time-buying" systematically after human data has not so far been addressed in the literature.

## 3 The System

Our system can bridge the gap between the user's request and the moment when it is ready to provide task information by producing similar utterances to those employed by humans in such situations. It simulates an automatic telephone system in a travel agency whose function is to receive requests for flights from customers and look for matching offers. Below is a description of the training process (Section 3.1) followed by an outline of the system architecture (Section 3.2).

### 3.1 Training a "Time-Buyer" Selection Strategy

#### 3.1.1 States and Actions

The possible actions for the system were taken from the "time-buying act classification scheme" proposed by López Gambino et al. (2017). This scheme includes 11 categories of utterances which humans employ in order to bridge time. However, we included only seven of these categories, due to several reasons. Utterances corresponding to the categories *filler* and *incomplete* were difficult to synthesize with the right prosody. Including category *confirmation/expansion request* would have introduced the risk of the user producing new content which we could not handle within our Wizard-of-Oz setup (see 4). Finally, we merged category *partial match* under *temporary non-availability*, since we did not find enough variety of non-availability utterances in the corpus and the functions of both categories are relatively similar.

Additionally, in order to further reduce the action and state spaces given the small size of the training data (see 3.1.3), we grouped these seven categories into two larger classes: *grounding* and *information state*. Table 2 lists the seven categories chosen and shows how they were grouped. On the other hand, we wanted our system to resemble, to some extent, human speakers' pausing behavior. Therefore, we explicitly included *pausing* in the action space. The resulting space thus consisted of four actions: **produce grounding action** and **produce information state utterances**, as in Table 2; and **pause for N seconds**, with $N = 2$ and $N = 4$ seconds.[2]

As for the state space, the state variables were the two last actions produced by the system: $a_{t-2}, a_{t-1}$. Given the four actions available, this resulted in 16 possible states.

[2]We originally chose 500 and 3000 ms as pause durations, following Jefferson (1989)'s suggestion of one second as the approximate maximum duration of an unmarked pause in conversation. However, we perceived the resulting production as sounding too rushed, which is why we extended pause durations to 2000 and 4000 ms.

### 3.1.2 Learning a Time-Buying Policy

We used OpenDial (Lison, 2015; Lison and Kennington, 2016) to estimate the probability and the utility of choosing each one of the available actions in each state.[3] OpenDial makes it possible to define a factored joint distribution (in the form of *probabilistic rules*), structured as sets of conditions together with the effects which may take place given those conditions. It is realised as a Partially Observable Markov Decision Process (POMDP) Bayesian Network model to estimate the distribution over all possible effects for each possible set of conditions. In our case, however, the model is best described as a simple Markov Decision Process (MDP), since the states are made up of the last two system actions (as explained in 3.1.1) and are thus fully observable.

There were 16 possible dialogue states and each state could result in four possible system actions. Input variables represent the rule conditions, i.e. the values of the state variables, and output variables represent the rule effects, namely the actions selected by the system (see Fig. 1). This resulted in 16 rules, one for each dialogue state. As an example, the rule corresponding to the state in which the last actions are *produce grounding utterance* and *produce long pause* respectively is structured as follows:

if $a_{t-2} == grounding$ and $a_{t-1} == long\ pause$:
decision = *grounding* (util = theta_grounding)
decision = *information state* (util = theta_state)
decision = *long pause* (util = theta_long_pause)
decision = *short pause* (util = theta_short_pause)

This rule shows the four possible values that the variable `decision` can take up, followed by the utilities corresponding to them. The four parameters starting with *theta* are the utility values which will be learned. The goal of training the system is to learn the probabilistic mapping from the input of the rule to its possible output values.

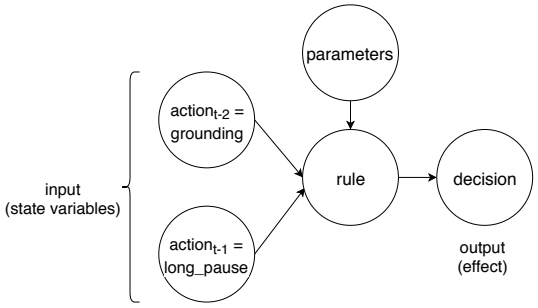


Figure 1: Example of Bayesian network connecting a rule with its input and effect

### 3.1.3 Data

The training data were extracted from the DSG-Travel Corpus (López Gambino et al., 2017). The corpus consists of 92 human-human dialogues which resulted from a role play activity simulating phone calls to a travel agency. One of the speakers plays the role of the caller, a customer who wants to buy a flight, whereas the other one acts as the travel agent who looks through a list for a matching flight to offer the caller. We only used the speech of the participant playing the travel agent, and specifically the parts of the dialogues between the customer's request and the information presentation stage, i.e. the period during which the travel agent buys time while looking for a matching flight. This resulted in 801 utterances.

OpenDial has provisions for using Wizard-of-Oz derived data for training. We were thus able to obtain 801 sequences of actions ($a_{t-2}$, $a_{t-1}$, $a_t$) representing the speaker's decision at time *t* and the two immediately previous decisions as part of the input state.

### 3.1.4 Parameter Estimation

The intial prior of the MDP was modeled with a Dirichlet distribution for probability rules and a Gaussian distribution for utility rules. OpenDial applies Bayesian learning to estimate the posterior distribution

[3] `http://www.opendial-toolkit.net`

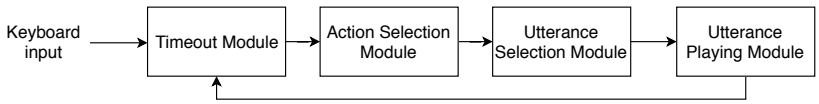


Figure 2: System architecture

$P(\theta|\mathcal{D})$, where $\mathcal{D}$ is the set of state-action pairs in the training data and $\theta$ represents the rule parameters. This distribution can be expressed as below (following Lison (2015)):

$$P(\theta|\mathcal{D}) = \eta P(\theta) \prod_{\langle \mathcal{B}_i, a_i \rangle \in \mathcal{D}} P(a_i|\mathcal{B}_i; \theta)$$

where $P(a_i|\mathcal{B}_i; \theta)$ is the probability of action $a_i$ being selected in the state $\mathcal{B}_i$ with rule parameters $\theta$, and $\eta$ is a normalization factor. Thus, at each iteration, the parameters are updated as follows:

$$P(\theta_{(i+1)}) = \eta P(\theta_{(i)}) P(a_i|\mathcal{B}_i; \theta_{(i)})$$

### 3.2 System Architecture

The system was developed using InproTK$_s$ (Kennington et al., 2014) and it consists of four modules, as illustrated in Fig. 2.[4]

**Timeout Module:** The Timeout module receives input, checks whether a result can already be presented or whether it is still too early, and forwards its decision to the Action Selection module. Ideally, the input received by the Timeout module would be the user's speech. However, the version for our current evaluation did not include a speech recognition or language understanding component: Instead, a confederate entered signals through the keyboard (as explained in Section 4).

**Action Selection Module:** This module selects one among the possible actions listed in Section 3.1.1. The selection is based on the learned policy explained in Section 3.1.2. This decision is then forwarded to the Utterance Selection Module. After the corresponding utterance has been played, the Action Selection module chooses a new time-buying action, and this process continues until the system can announce a matching flight.

**Utterance Selection Module** This module has two main functions. The first one is choosing a time-buying category based on the decision received from the Action Selection module. For example, if the decision received is *grounding*, it will choose between *acknowledgment, commitment* and *echoing*; otherwise, if the decision received is *information state*, the choice will be between *agent/system state, availability, temporary non-availability* and *wait request* (See Table 2).

In order to make this selection, the module considers the frequency distribution, in the human-human data, of the available time-buying categories in the corresponding position in the interaction. For instance, if the decision received from the Action Selection module is *grounding* and the system has already produced two time-buying utterances, it will consider all the *grounding* utterances which appear in the data in the third position of the time-buying phase, together with their respective categories. Since the distribution for this position is *acknowledgment: 0.05, commitment: 0.28, echoing: 0.67*, the Utterance Selection module will sample from this distribution in order to select the next category. Due to the reduced size of the corpus, only the frequencies of the first six positions are considered: Starting from the seventh utterance, the module alternates between the probabilities for the fifth and sixth slots.

Once a category has been selected, the second task of the Utterance Selection module is to choose a specific utterance to send to the Utterance Playing module. Four utterances are available for each category. The decision is simply, out of these four utterances, the first one which has not been used yet (if all four have been used, the selection starts again at the beginning of the list). Finally, the utterance is forwarded to the Utterance Playing module, which plays an audio file with the synthesized utterance. On the other hand, if the decision received from the Action Selection module is not an utterance but a pause,

[4]InproTKs was taken from `https://bitbucket.org/inpro/inprotk`.

**PARTICIPANT:** **Ich würde gerne von Frankfurt nach Sydney fliegen, am 3. August und zwar vormittags.**
I'd like to fly from Frankfurt to Sydney, on August 3 in the morning.

**SYSTEM (FIXED):**

| **Mm-hm** | **einen kleinen Moment** | **warten Sie bitte noch einen Augenblick** | **Sekunde noch** | **Augenblick, bitte** | **Ich habe einen passenden Flug gefunden...** |
|---|---|---|---|---|---|
| Mm-hm | one moment, please | please wait a little longer | one more second | one moment, please | I have found a matching flight. |

**SYSTEM (RANDOM):**

| **Mm-hm** | **vormittags** | **da haben wir was im Angebot** | **da gucken wir doch mal** | **okay** | **schaue ich gerade einmal nach** | **Ich habe einen passenden Flug gefunden...** |
|---|---|---|---|---|---|---|
| Mm-hm | in the morning | we have something to offer you | let's see | okay | I'm having a look | I have found a matching flight. |

**SYSTEM (LEARNED):**

| **Mm-hm** | **einen kleinen Moment, bitte** | **nach Sydney** | **am 3. August** | **Sekunde noch** | **ich schaue gerade mal in meine Liste** | **Ich habe einen passenden Flug gefunden...** |
|---|---|---|---|---|---|---|
| Mm-hm | one moment, please | to Sydney | on August 3 | one more second | I'm having a look in my list | I have found a matching flight. |

Figure 3: Examples of the three time-buying strategies employed by the system (original utterances in German in bold; English translation provided below). The gray intervals represent pauses.

the task of the Utterance Selection module is simply limited to forwarding this decision to the Utterance Playing module. Utterances were synthesized with Cereproc's male voice for German, "Alex".[5]

In summary, to make a decision about a particular act at a given point, the system first checks whether information can already be presented (Timeout module); if not, it selects a high-level act based on the learned policy (Action Selection module) and, based on that, an actual utterance (Utterance Selection module), which it then realizes (Utterance Playing module). The division of the decision-making process between the Action Selection and Utterance Selection modules was due to the reduced size of the training data: Grouping all utterances into two broad categories in the Action Selection module (*grounding* and *information state*) and refining the decision in the Utterance Selection module made it possible to keep the state space smaller for learning the parameters of the action selection rules (see 3.1).

## 4 Evaluation: Comparing Learned, Random, and Rule Based Time-Buying

In order to evaluate generation output without having a full dialogue system, we integrated the system into a Wizard-of-Oz environment. The Wizard's task was to press a key whenever she judged that the participant's request was complete. The system then acknowledged the request by producing *mm-hm* and started to buy time. The Wizard could also trigger clarification requests when the participant forgot to mention one of the search criteria or the request had not been expressed clearly; examples of these are *Could you please repeat the destination airport?* and *Do you prefer a specific airline?*. Participants were told that they interacted with a fully automated system. Below we provide more details about the experimental design and procedure of the evaluation, as well as the participants.

**Design:** There were three experimental conditions: LEARNED, RANDOM and FIXED. The difference between the conditions was the strategy used by the system to bridge the gap between the user's request and the moment when it announces finding a flight (see Fig. 3 for examples):

**FIXED** : The system bridges the gap by explicitly asking the user to wait, through utterances such as *please wait; one moment, please; give me a second,* etc. The utterances are separated by four-second intervals.

**RANDOM** : The system bridges the gap by randomly selecting from a set of utterances similar to those found in the DSG-Corpus (see 3.1.3). In between utterances, the system can also randomly choose to produce a four-second pause, a two-second pause or no pause at all.

**LEARNED** : The system employs the learned strategy described in 3.1.2. The utterances are the same as in the RANDOM strategy.

Participants were presented with each one of these conditions four times, in random order.

[5] https://www.cereproc.com/

| St. | FIXED (total sum) | FIXED (median) | FIXED (iqr) | RANDOM (total sum) | RANDOM (median) | RANDOM (iqr) | LEARNED (total sum) | LEARNED (median) | LEARNED (iqr) |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 427 | 4 *** | 1 | 452 | 4 * | 1 | 486 | 4 | 1 |
| 2 | 460 | 4 ** | 2 | 471 | 4 | 2 | 496 | 4 | 1 |
| 3 | 376 | 3 *** | 1 | 402 | 3 *** | 1 | 456 | 4 | 2 |

Figure 4: Ratings received by each strategy, by statement: 1) *It was pleasant to interact with this system*, 2) *The system provided an answer within an appropriate amount of time*, 3) *The system acts the way I would expect a person to act. iqr* stands for interquartile range. (* $p<.017$, ** $p<.003$, *** $p<.0003$)

**Procedure** Each participant played the role of a secretary at a company, who had been instructed to call a travel agency to book a number of flights for some of the company executives. Participants were told that they would be speaking to an automatic system which could understand speech, and they received a handout with a list of items. Each item contained the criteria defining a flight that the participant should request, e.g. *Frankfurt-Sydney, May 24, Lufthansa*. The calls started with the system greeting the participant. After this greeting, the participant asked for one of the flights on the list. Following this request, the Wizard pressed a key for the system to produce *mm-hm* in order to signal having received the participant's request and subsequently start buying time. After 20 seconds, the system announced having found a flight and told the participant that the flight details would be sent to the company by email.[6] We chose 20 seconds as the duration of the time-buying stretch because this is close to the average duration of the time-buying stretches in the human-human corpus (17.5 seconds). Finally, if the participant said "goodbye", the Wizard pressed a key for the system to say "goodbye" as well.

After every call, participants were given some time to rate the system. For each of the statements below, they chose an option from 1 (completely disagree) to 5 (completely agree):[7]

- It was pleasant to interact with this system.
- The system provided an answer within an appropriate amount of time.
- The system acts the way I would expect a person to act.

There was also an optional field for further comments. Once the participant had completed the assessment, the next call started, with the system greeting the customer as before.

Each participant completed 14 calls: two test calls for making sure they had understood the instructions and 12 experiment calls. Participants were instructed to include only one flight per call.

**Participants:** Thirty participants were involved in the study, 19 female and 11 male, recruited through flyers left at the University cafeteria, by email or on the Facebook group of the University.

**Analysis:** We compared the ratings given to each of the strategies (FIXED, RANDOM and LEARNED) for each of the three statements rated (see 4) and tested significance of differences through Wilcoxon signed-rank test. We also applied Bonferroni correction due to the multiplicity of statements per stimulus, which resulted in the following significance levels: 0.05/3 = **.017**; 0.01/3 = **.003**; 0.001/3 = **.0003**.

### 4.1 Results

No significant differences were found between the FIXED and RANDOM strategies. In contrast, the LEARNED strategy was rated significantly better than the FIXED strategy for all three statements ($Z$=356, $p<.0003$; $Z$=475, $p<.003$ and $Z$=800, $p<.0003$). Additionally, LEARNED was rated significantly better than RANDOM for statements 1 ($Z$=652, $p<.017$) and 3 ($Z$=904, $p<.0003$). Fig. 4 shows the total sum of the ratings assigned to each condition in each statement, the median score and the interquartile range.

[6] We told participants that the system already had the contact details of the company, the latter being a frequent customer

[7] The questionnaire was adapted from (López Gambino et al., 2018).

## 5 Discussion and future work

It has been claimed that systems which bridge time through speech are preferred by humans over those which wait for the information in silence (Tom et al., 1997; López Gambino et al., 2018). In this experiment, we tested three time-bridging strategies involving speech, with a view to identifying the characteristics that this speech must have in order to render the interaction natural and pleasant for users. In particular, we focused on two aspects: *variety* and *sequencing*. In the FIXED condition, no attention is paid to either of these aspects, since all utterances realize the same dialogue act, namely requesting extra time, and they are presented in random order. The RANDOM condition includes a variety of utterances representing different dialogue acts, but the way in which they are presented is also random. Finally, the LEARNED condition considers knowledge about both the variety observed in humans' time-buying strategies and their distribution with respect to the moves preceding them and to their position in the time-buying stretch. Therefore, our results suggest that both aspects —variety and sequencing— play a role in shaping user experience, since our model received higher ratings than the other two strategies.

On the other hand, it is worth mentioning that the LEARNED system was rated as capable of finding a result in a more appropriate amount of time than the FIXED system, even though waiting time was 20 seconds for every dialogue, regardless of the strategy employed. What is yet more interesting is that this difference was not found between the LEARNED and RANDOM conditions. A possible hypothesis is that repetition of the same dialogue act in the FIXED condition might have led to users' annoyance and, consequently, to a perception of waiting time as longer, something which did not happen in the other two conditions, in which moves were more varied and potentially more "entertaining".

It must be noted that, although we trained an MDP model, the probability function resulting from the learning process was near-uniform and the system's choices were thus controlled by the utility function. Therefore, the model actually learned resembles a trigram model, since the system always selects the action with the highest utility given the two previous actions. An MDP model might prove useful in future work in which other variables —such as the user's speech— are considered.

A further consideration worth addressing is the issue of human-likeness. The importance of human-likeness for dialogue systems has been discussed at length (Turing, 1950; Reichman, 1985; Dahlbäck et al., 1993; Larsson, 2005; Edlund et al., 2008; Baumann and Schlangen, 2013; Traum, 2018). Although it seems clear that there are aspects deserving a higher priority —such as clarity— in our results, human-likeness of the strategy used correlates with reported pleasantness of interaction, suggesting that users are not impervious to this characteristic.

In addition, an aspect which is not addressed in this study but certainly deserves further research is the relation between time bridging and estimated time until task content is available or, in other words, whether the characteristics of the speech used to buy time are somehow influenced by the predicted length of the information delay.

In the future, we want to endow the system with the ability to interact with the user while buying time, making it more conversational and responsive. We would also like the system to be able to make incremental decisions based on both user speech and the amount and quality of the task information available to it, so that it can leverage this information while buying time. This would, for example, result in behavior such as *Unfortunately I can't see any flights in the morning, I only have... oh, one second, I just found one flight in the mor-, actually two flights in the morning...*. Finally, it would be interesting to analyze the characteristics of time-buying in corpora corresponding to other domains, and also perhaps in other languages.

## 6 Acknowledgments

This research/work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG). Special thanks to Ayten Tüfekci for her valuable help carrying out experiments, as well as to Nikolai Ilinykh, Nazia Attari and Ting Han for interesting discussions about time-buying.

## References


Gerrit Antonides, Peter Verhoef, and Marcel van Aalst. 2002. Consumer perception and evaluation of waiting time: A field experiment. In *Journal of Consumer Psychology*, volume 12 (3), pages 193–202. Lawrence Erlbaum Associates, Inc.

Timo Baumann and David Schlangen. 2013. Open-ended, extensible system utterances are preferred, even if they require filled pauses. In *Proceedings of Short Papers at SIGdial 2013*.

S. Betz, B. Carlmeyer, P. Wagner, and B. Wrede. 2017. Interactive hesitation synthesis and its evaluation. preprint at `https://www.preprints.org/manuscript/201712.0058/v1`.

H. Buschmeier, T. Baumann, B. Dosch, S. Kopp, and D. Schlangen. 2012. Combining incremental language generation and incremental speech synthesis for adaptive information presentation. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 295–303.

Okko Buß and David Schlangen. 2010. Modelling sub-utterance phenomena in spoken dialogue systems. In *Proceedings of the 14th International Workshop on the Semantics and Pragmatics of Dialogue (Pozdial 2010)*, pages 33–41, Poznan, Poland, June.

Herbert Clark and Jean Fox Tree. 2002. Using uh and um in spontaneous speaking. In *Cognition*, volume 84 (1), pages 73–111. Elsevier Science.

Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of oz studies: Why and how. *Knowledge-Based Systems*, 6(4):258 – 266.

Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. In *Speech Communication*, volume 50, pages 630–645. Elsevier.

I. Hirsch, R. Bilger, and B. Heatherage. 1950. The effect of auditory and visual background on apparent duration. In *American Journal of Psychology*, volume 69. University of Illinois Press.

Gail Jefferson. 1989. Notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In D. Roger and P. Bull, editors, *Conversation: An interdisciplinary perspective*. Multilingual Matters, Clevedon, UK.

Casey Kennington, Spyros Kousidis, and David Schlangen. 2014. InproTKs: A toolkit for incremental situated processing. In *Proceedings of SigDial*, pages 84–88, Philadelphia, USA. ACL.

Lea Susan Kohtz and Oliver Niebuhr. 2017. How long is too long? how pause features after requests affect the perceived willingness of affirmative answers. In *Proceedings of the International Conference on Spoken Language Processing*.

Staffan Larsson. 2005. Dialogue systems: Simulations or interfaces? In *Proceedings of the Ninth Workshop on the Semantics and Pragmatics of Dialogue*.

Steven Levinson. 1983. *Pragmatics*. Cambridge University Press, Cambridge, England.

P. Lison and C. Kennington. 2016. OpenDial: A toolkit for developing spoken dialogue systems with probabilistic rules. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - System Demonstrations*.

Pierre Lison. 2015. A hybrid approach to dialogue management based on probabilistic rules. *Computer Speech & Language*, 34(1):232 – 255.

Soledad López Gambino, Sina Zarrieß, and David Schlangen. 2017. Beyond on-hold messages: Conversational time-buying in task-oriented dialogue. In *Proceedings of SIGdial 2017*.

Soledad López Gambino, Sina Zarrieß, and David Schlangen. 2018. Testing Strategies For Bridging Time-To-Content In Spoken Dialogue Systems. In *Proceedings of the Ninth International Workshop on Spoken Dialogue Systems Technology (IWSDS 2018)*.

Kristina Lundholm Fors. 2015. *Production and Perception of Pauses in Speech*. Ph.D. thesis, University of Gothenburg.

Nira Munichor and Anat Rafaeli. 2007. Numbers or apologies? Customer reactions to telephone waiting time fillers. In *Journal of Applied Psychology*, volume 92 (2), pages 511–518. American Psychological Association.

Rachel Reichman. 1985. *Getting Computer to Talk Like You and Me*. The Massachusetts Institute of Technology, Cambridge, Massachusetts.

D. Schlangen and G. Skantze. 2011. A general, abstract model of incremental dialogue processing. In *Dialogue and Discourse*, volume 2 (1), pages 83–111.

Gabriel Skantze and Anna Hjalmarsson. 2010. Towards incremental speech generation in dialogue systems. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Amanda Stent. 1999. Content planning and generation in continuous-speech spoken dialog systems. In *Proceedings of the KI'99 workshop "May I Speak Freely?"*.

Gail Tom, Michael Burns, and Yvette Zeng. 1997. Your life on hold: The effect of telephone waiting time on customer perception. In *Journal of Direct Marketing*, volume 11 (3), pages 25–31. John Wiley and Sons, Inc. and Direct Marketing Educational Foundation, Inc.

David Traum. 2018. Beyond dialogue system dichotomies: Principles for human-like dialogue. Presentation – International Workshop on Spoken Dialogue Systems.

Vivian Tsai, Timo Baumann, Florian Pecune, and Justine Casell. 2018. Faster responses are better responses: Introducing incrementality into sociable virtual personal assistants. In *Proceedings of the Ninth International Workshop on Spoken Dialogue Systems Technology (IWSDS 2018)*, Singapore.

Alan Turing. 1950. Computing machinery and intelligence. *Mind*, 59:433–460.